EHWGesture - A Dataset for Multimodal Understanding of Clinical Gestures

EHWGesture - A dataset for multimodal understanding of clinical gestures

Gianluca Amprimo[1*] Alberto Ancilotto[2] Alessandro Savino[1] Fabio Quazzolo[1]
Claudia Ferraris [3] Gabriella Olmo [1] Elisabetta Farella [2] Stefano Di Carlo [1]
[1]Department of Control and Computer Engineering, Politecnico di Torino, Torino, Italy
[2] Fondazione Bruno Kessler, Trento, Italy
[3]CNR-IEIIT, Torino, Italy
name.surname@polito.it, {aancilotto, efarella}@fbk.eu, claudia.ferraris@cnr.it

Abstract

Hand gesture understanding is essential for several applications in human-computer interaction, including automatic clinical assessment of hand dexterity. While deep learning has advanced static gesture recognition, dynamic gesture understanding remains challenging due to complex spatiotemporal variations. Moreover, existing datasets often lack multimodal and multi-view diversity, precise ground-truth tracking, and an action quality component embedded within gestures. This paper introduces EHWGesture, a multimodal video dataset for gesture understanding featuring five clinically relevant gestures. It includes over 1,100 recordings (∼6 hours), captured from 25 healthy subjects using two high-resolution RGB-Depth cameras and an event camera. A motion capture system provides precise ground-truth hand landmark tracking, and all devices are spatially calibrated and synchronized to ensure cross-modal alignment. Moreover, to embed an action quality task within gesture understanding, collected recordings are organized in classes of execution speed that mirror clinical evaluations of hand dexterity. Baseline experiments highlight the dataset's potential for gesture classification, gesture trigger detection, and action quality assessment. Thus, EHWGesture can serve as a comprehensive benchmark for advancing multimodal clinical gesture understanding.

1. Introduction

Automatic hand gesture recognition is a complex yet significant challenge. Gestures are fundamental to human interaction and, along with facial expressions, constitute the non-verbal component of inter-subject communication [19].

Computer systems capable of recognizing hand gestures can support various essential tasks. For instance, conventional applications include sign language recognition [25], augmented reality [30], robotics, and human-computer interaction [15]. Early gesture recognition methods primarily relied on wearable devices, such as gloves embedded with inertial measurement units [24]. Nowadays, computer vision methods, especially those based on deep learning [14], have become state-of-the-art, offering a non-invasive and more natural interaction.

Gestures can be categorized as static or dynamic [26]. Static hand poses are analyzed in still images, without any temporal evolution of the hand position. Large datasets, especially those based solely on RGB data, can be easily crowdsourced as they require minimal instrumentation (*e.g.*, commercial webcams), as demonstrated by the HaGRID dataset by Kapitanov *et al*. [15]. Also researchers in dynamic hand gesture understanding have leveraged crowdsourcing to collect large-scale datasets, as done by Materzynska *et al*. for the Jester dataset [18]. However, this approach does not allow to capture multimodal data, despite the potential benefits of incorporating additional modalities such as depth or event-based data to handle the complex spatial and temporal variations typical of dynamic gestures.

In particular, event-based gesture datasets that combine neuromorphic data with other modalities remain scarce, making cross-modality comparisons challenging. Furthermore, many dynamic gesture datasets rely solely on automated annotation using deep-learning-based hand tracking methods like MediaPipe [21]. While efficient, this may result in the absence of a strong ground truth, particularly for precise temporal segmentation of gestures. For example, tracking joint positions with a motion capture system would provide more accurate validation and enhance the reliability of gesture understanding benchmarks.

Finally, dynamic gesture recognition can be extended to gesture *understanding*, which involves not only classifying gestures within a given time window but also segmenting their *triggering* phases (*e.g.*, the tapping phase in finger pinching) [21] or assessing the quality of execution of the gesture (*e.g.*, recognizing gesture speed) [9]. These

*Corresponding author.

Figure 1. Dynamic gestures included in the dataset and keyframes acquired by the event (center) and RGB-D (left and right) cameras. From top-left: Finger tapping, hand opening and closing, finger-to-nose, and pronation-supination gestures.

tasks are central in innovative applications such as automatic hand dexterity assessment of clinical gestures. Indeed, a growing trend at the intersection of medicine and computer vision explores hand tracking and gesture recognition to automatically assess motor impairments caused by conditions such as Parkinson's disease [3]. This research typically involves multiple stages, including hand detection and tracking in video, gesture recognition to identify specific movements, and action quality assessment (AQA) to quantify performance based on clinical ratings of disease severity [9]. However, large-scale, publicly available benchmarks in this domain are lacking [3]. While pathological data are crucial for studying specific disease-related impairments, even extensive datasets from non-pathological subjects could improve the robustness of early-stage gesture recognition and tracking, which continue to face significant accuracy challenges.

This work introduces EHWGesture, a large-scale multimodal video dataset for dynamic gesture understanding. The dataset comprises five gestures commonly used in clinical hand dexterity assessments, performed over 1,100 trials, each lasting 20s, for a total of more than six hours of recordings. Gestures were captured using two high-resolution RGB-Depth cameras positioned at orthogonal viewpoints, recording at 30 frame per seconds (fps), resulting in more than 2.6M RGB and depth frames. Additionally, a neuromorphic event camera sampled events at a high frequency (100 MHz). All recordings took place within the tracking volume of a motion capture system, which tracked key hand landmarks involved in the gestures. This *gold standard* system provides precise ground truth labeling and enables validation against a highly accurate reference. All recording streams were synchronized to maintain temporal alignment across modalities. Spatial calibration data are also provided to facilitate the fusion of different data modalities from a spatial perspective. Finally, some gestures were performed at varying speeds, with volunteers following a metronome. This enables AQA based on execution speed, adding a new dimension to gesture understanding that mimics clinical differences in hand dexterity.

This work also presents a series of experiments using the newly collected dataset to establish baseline models for the main addressed tasks. To summarize, its key contributions are:

- Providing a large-scale multimodal dataset, integrating synchronized and spatially calibrated RGB, depth, and event-based data.
- Offering high-quality ground truth via a motion capture system for precise hand landmark tracking, enabling accurate validation and benchmarking.
- Introducing controlled-speed recordings guided by a metronome, allowing for AQA of the recorded gestures based on execution speed, similarly to clinical assessments.
- Establishing baseline models on the EHWGesture dataset for gesture classification, temporal segmentation (*i.e.*, triggering detection), and AQA.
- Investigating the impact of different modalities, input sequence length and frame rate on gesture understanding.

2. Related Works

Dynamic hand gesture datasets in the literature focus on different tasks depending on their applications (*i.e.*, sign language [6], conversational gestures, and interaction gestures). Large scale and open datasets including several clinical hand gestures are currently lacking, as most of the experiments rely on limited and private datasets [3]. The PD4T dataset [9] is the only exception, with data collected from 30 patients with Parkinson's disease to perform AQA. Despite

Table 1. Comparison of dynamic hand gesture datasets. Acronyms: RGB – Red, Green, Blue, D - Depth, E - Event, GR – Gesture Recognition, GD – Gesture Detection, GT – Gesture Triggering, AQA – Action Quality Assessment, HT – Hand Tracking, HGT - Hand Ground-Truth, SBJ -Subjects, PoV - Point of View, IR - Infra-Red

Dataset	Videos	SBJ	Classes	Frames	Modality & Resolution (px)	HGT	Tasks	PoV	Views
PD4T [9]	1,654	30	2	-	RGB: 856×480	-	GR, AQA	3rd	1
EgoGesture [30]	2,081	50	83	2,953,224	RGB, D: 640×480	-	GR, GD	1st	1
FHANDS [10]	1,175	6	27	105,459	RGB: 1920×1080, D: 640×480	✓	GR, HT	1st	1
IPN Hand [5]	200	50	13	800,000	RGB: 640×480	-	GR, GD	3rd	1
ChaLearnConGD [27]	22,535	21	249	-	RGB, D: 240×320	-	GR, GD	3rd	1
ChaLearnIsoGD [27]	47,933	21	249	-	RGB, D: 240×320	-	GR	3rd	1
Jester [18]	148,092	1,376	27	5,331,312	RGB: 100×variable	-	GR	3rd	1
NVGesture [20]	1,532	20	25	-	RGB, D, IR: 320×240	-	GR, GD	3rd	2
DVSGesture [2]	1,342	29	11	-	E: 128×128	-	GR, GD	3rd	1
NavGesture [17]	1,681	28	6	-	E: 640×480	-	GR, AQA	3rd	1
EB-HandGesture [1]	9,000	5	6	-	E: 640×480	-	GR, AQA	3rd	1
EHWGesture	**3,300***	**25**	**11**	**3,300,000****	**RGB: 1920×1080, D: 640×480, E: 320×240**	✓	**GR, GT, GD, AQA**	**3rd**	**3**

* 1,100 recordings x (2 RGB-D cameras + 1 event camera) ** RGB + D + E frames, E accumulated using a 33 ms window

containing more then 1600 recordings, the dataset is limited to only two classes of clinical hand gestures (finger tapping, hand opening-closing) and one single modality (RGB).

For a more comprehensive overview, this section broadens the perspective to gesture understanding for interaction, as these gestures are the most similar to those included in the EHWGesture dataset. Previous datasets aimed to address specific limitations in the field, such as small sample sizes—particularly in terms of subjects—, limited background diversity, restricted data modalities, and varying annotation types. Table 1 provides a structured comparison of the most relevant datasets in this domain, alongside the EHWGesture dataset.

Most datasets provide recordings from a third-person perspective, except for EgoGesture [30] and FHANDS [10], which are specifically designed for egocentric gesture recognition. All datasets except NVGesture and the proposed EHWGesture capture recordings from a single viewpoint [20]. The largest dataset in terms of recorded frames and subjects, Jester [18], contains only RGB recordings, as it was obtained through crowdsourcing. EgoGesture, FHANDS, ChaLearn ConGD, and IsoGD [27] provide both RGB and depth data, whereas NVGesture also includes stereoscopic infrared recordings. For the latter, additional modalities, such as optical flow and segmentation masks, were derived from RGB, but Table 1 considers only raw acquired modalities. Most datasets are annotated for gesture recognition using manually segmented gestures. However, IPN Hand [5], DVSGesture [2], ChaLearn ConGD, and NVGesture also include annotations for continuous gesture detection, *i.e.*, identifying gestures within an unsegmented recording that includes both gesture and non-gesture segments. FHANDS [10] also provides ground-truth hand tracking data for key hand landmarks, obtained using a wearable system based on magnetic sensors. As a result, this dataset can also be used for hand tracking tasks in the context of hand-object interaction. Datasets for event-based gesture recognition are less common, as neuromorphic cameras are still emerging. In event cameras, light reaching a pixel is converted into a voltage. Any deviation from a reference voltage is detected, and when it exceeds a predefined threshold, an event is triggered. These events are transmitted with extremely high temporal resolution (*e.g.*, 100 MHz), and their data rate varies depending on the frequency of illumination changes in the scene.

The three event-based datasets in Table 1 (DVSGesture [2], EB-HandGesture [1], and NavGesture [17]) do not provide complementary modalities for direct comparisons between neuromorphic vision and conventional approaches. DVSGesture was one of the first large-scale event-based gesture datasets and has been widely used for both isolated gesture recognition and continuous gesture detection [2, 11, 31]. NavGesture examined gesture detection in static versus dynamic conditions, such as while walking, by combining an event camera with a smartphone. EB-HandGesture supports both gesture recognition and AQA by recording gestures at three different speeds, though this aspect was not explored by the authors when introducing the dataset [1].

EHWGesture addresses current gaps by:

- Providing the first dataset including multiple gesture from clinical assessment of hand dexterity, with timed gesture execution, mimicking real patients conditions, to incorporate AQA into gesture recognition.

- Including simultaneous and calibrated recordings from event, RGB and depth cameras from three different viewpoints, for cross-view and cross-modality comparisons
- Offering ground-truth hand tracking through a high-precision motion capture system, ensuring fast and accurate temporal segmentation and spatial reference for key hand landmarks involved in the gesture.

The next section details the dataset and the collection pipeline used.

3. EHWGesture Dataset

EHWGesture was developed as a large-scale, multimodal dataset with high-quality ground truth annotations to support the recognition and characterization of five gestures commonly assessed in clinical evaluations. These gestures are based on the Unified Parkinson's Disease Rating Scale (UPDRS), a widely used framework for evaluating hand dexterity in Parkinson's disease [12]. Four of these movements—finger tapping, hand opening and closing, hand pronation-supination, and finger-to-nose reaching— are dynamic, while one static gesture involves extending the arm forward to assess distal tremors.

The five gestures, captured across multiple modalities and viewpoints, are illustrated in Figure 1. Moreover, previous studies have shown that pre-training on other AQA tasks, even from non-clinical domains, can aid Parkinson's video staging [9]. For this reason, the protocol was structured so that the finger tapping, pronation-supination, and opening-closing gestures were also performed at different execution speeds, embedding an AQA task focused on pace recognition. Slowness in movement (*bradykinesia*) is a hallmark of Parkinson's disease. Thus, execution speed is typically linked with AQA for Parkinson's disease detection [23], further reinforcing the decision of embedding this kind of analysis in the dataset. This aspect could be leveraged by future clinical studies to support pretraining of deep learning models for clinical AQA despite the current paucity of data from real patients.

3.1. Dataset creation

The dataset was created using video recordings from 25 healthy volunteers, all of whom provided written informed consent for data collection and sharing. Participants ranged in age from 24 to 65 years, with 7 female and 18 male subjects. The multimodal nature of the dataset imposed constraints on the sample size and recording conditions. Nonetheless, efforts were made to include participants with diverse hand shapes and skin tones.

Instrumentation: Recordings were conducted in a laboratory equipped with an Optitrack (OPT) system featuring six Prime13 cameras (1280×1024px resolution). OPT cameras operated at 120 Hz, covering a working volume of approximately 6×4×3 m^3. Passive reflective markers from the OPT system appeared in RGB frames and could also interfere with depth data by creating holes in the depth map. To minimize these effects, small reflective markers and a minimal marker schema were used, tracking only key hand landmarks relevant to each gesture rather than performing full hand tracking. Two RGB-Depth cameras (Microsoft Azure Kinect) were placed within the motion capture volume, positioned orthogonally so that their viewpoints intersected near the center of the capture area. Each Kinect device recorded RGB and depth frames at 30 fps, with resolutions of 1920×1080px and 640×480px, respectively. Between the two devices, an Inivation DVXplorer Lite event camera captured frontal recordings of the gestures, operating at 100 MHz with a resolution of 320×420px.

Synchronization and calibration: Temporal alignment of all recordings was achieved using an OptiTrack eSync2. The motion capture system coordinated all devices, managing acquisition start times and synchronizing exposure intervals for the Kinect cameras. The event camera operated independently, but a trigger event was generated each time an RGB-Depth frame was captured. This enabled the identification of events within the RGB frame acquisition, allowing event data to be accumulated into frames closely aligned with RGB images. Each device underwent individual calibration using appropriate procedures. Additionally, stereo calibration was performed by capturing a checkerboard pattern displayed on a PC screen, with three OPT markers placed at specific corners. The PC screen was required as pixel refreshing generated clear events that were also detected by the DVXplorer Lite camera. Recordings were taken from 40 different positions within the tracking volume, corresponding to likely hand locations during gesture execution. This extrinsic calibration is included in the dataset to support cross-device alignment.

Acquisition protocol: Each subject sat comfortably in the center of the acquisition space. Full-body recordings were captured while subjects performed the gestures. Each subject executed the five tasks, first with one hand and then the other. Each task was continuously recorded for 20 seconds, mirroring clinical assessments where gestures are evaluated over time rather than in isolation. Moreover, longer videos can be segmented into different window sizes, introducing greater internal variability compared to isolated gestures. Finger tapping, hand opening and closing, and pronation-supination were timed using a metronome at three speeds: SLOW (75 bpm), NORMAL (115 bpm), and FAST (145 bpm). Subjects were instructed to follow the metronome as closely as possible. Executions excessively deviating from the request were discarded during acquisition and repeated again. Each task was recorded twice for data augmentation.

Annotation: Additional data were collected, including hand shape measurements (length and width) and illumina-

tion conditions, as recordings spanned multiple days. Each recording was labeled with the corresponding gesture name and, for timed gestures, the execution speed. Furthermore, motion capture data provided reference points for temporal segmentation, as continuous gesture repetitions generated periodic patterns in key hand landmarks. These patterns were segmented automatically or semi-automatically by identifying extreme points in the signals, allowing large datasets to be efficiently labeled with a strong ground truth. To demonstrate the utility of motion capture data for this purpose, *triggers* were automatically extracted, representing key moments within gestures. For example, in finger tapping, triggers corresponded to pinching instants; in hand opening and closing, the fully closed hand; in finger-to-nose reaching, the outward-reaching hand; and in pronation-supination, the moments when the palm or back of the hand faced the cameras.

3.2. Dataset characteristics

A total of 1,100 recordings were conducted (44 per subject, 22 per body side), resulting in 2,200 RGB-Depth videos and 1,100 event recordings. The synchronized RGB (1920×1080px) and depth (640×480px) frames across both cameras amount to 2,640,000 frames. While events are collected as sequential data on a spatial grid of 320×240px, aggregating them into frames at 30 fps (aligned with the RGB-D cameras) generates an additional 660,000 frames, bringing the total to 3,300,300 frames. These statistics refer only to raw data; additional derived modalities, such as 3D point clouds, optical flow, or segmentation masks, could be extracted from the depth and RGB data streams.

The recorded hands were positioned between 30 and 100 cm from each camera, as subjects sat in a fixed location but adjusted their posture for comfort. Although gesture timing was enforced for some tasks, spatial amplitude was not restrained. Furthermore, the extended duration of each recording (20s) inherently introduced variability.

Considering the combined gesture recognition and AQA tasks, recordings may be grouped in 11 classes, as reported in Table 1 (two free gestures plus three gestures $\times$ three velocities). Moreover, while we do not provide hand bounding boxes or full hand tracking models, this information can be easily obtained using tools such as `mediapipe-hands` [29], as applied in our preprocessing (see supplementary materials). Additionally, motion capture data can be used to validate results from pre-trained 3D hand tracking networks, as demonstrated in [4]. Finally, given the available depth data and motion capture annotations, the dataset could also support the development of new multimodal 3D hand tracking models, particularly in pre-training phases.

4. Experiments

We showcase the potential of the EHWGesture dataset for gesture understanding through two experimental use cases and their respective baseline models.

First, we train models for gesture classification and AQA. These tasks are key to advancing gesture recognition systems for clinical assessment and quality-based analysis. This experiments employ unsegmented recording windows, making it more challenging than isolated gesture recognition, as windows may contain incomplete gestures. For gesture recognition, the five gestures, ignoring velocity classes, are classified. For AQA, the objective is to classify gestures based on the three execution speed classes (SLOW, NORMAL, FAST).

The second experiment focuses on detecting trigger events within recordings, using motion capture data as ground truth. This demonstrates the applicability of EHWGesture also to other field than clinical assessment, such interactive gesture research, and highlights the advantages of motion capture for precise annotation.

The data and the code used for these experiments are available at https://github.com/smilies-polito/EHWGesture.

4.1. Architectures

The dataset was benchmarked using three different convolutional architectures: PhiNet-3D [22], 3D ResNet-50 [13], and 3D ResNeXt-152 [28]. The 3D architectures were derived from their respective 2D versions following the approach in [16]. These models cover a broad range of computational complexities and network sizes, from the smallest PhiNet with $4.9M$ parameters to the largest ResNeXt with $116M$ parameters.

The two addressed tasks—gesture recognition and AQA—require different types of information. Gesture recognition primarily relies on spatial features, as shown in previous studies that solved the task using single-frame models [15]. In contrast, AQA depends on temporal information, requiring models that effectively capture motion dynamics [17].

To process the multimodal data, the networks were used to extract features from different input streams. A late feature fusion strategy, similar to [7], was then applied to classify the multimodal input, as illustrated in Figure 2. This approach enables the evaluation of how integrating RGB, depth, and event-based representations separately contributes to the gesture understanding task.

4.2. Training Setup

The training data consists of cropped RGB, depth, and event-based frames processed as described in the supplementary materials. All input frames were resized to a uniform resolution of 240p to ensure consistency across modalities. We defined a train/validation split that reserves data

from five subjects for evaluation. Networks were trained using the `SGD` optimizer with an initial learning rate of 10^{-3}, which was reduced by a factor of 0.1 every 10 training epochs.

Multimodal Contrastive Pretraining: Before finetuning for classification, each network underwent a multimodal contrastive pretraining phase. The SimCLR [8] framework was adapted for contrastive pretraining with multimodal inputs. In particular, to accommodate the dataset's multi-camera, multimodal nature, the original data augmentation strategy was modified. Specifically, given a sequence of frames $x_t^{(c,m)}$ at timestamp t, with modality m and camera c, positive and negative pairs were defined as follows:

- Positive pairs: frame sequences with the same timestamp but from different cameras or modalities:

$$P : ({x_{t1}}^{(c_1,m_1)}; {x_{t2}}^{(c_2,m_2)}) \text{ with } t_1 = t_2 \\ \text{and either } c_1 \neq c_2 \text{ or } m_1 \neq m_2. \quad (1)$$

- Negative pairs: frame sequences starting from different timestamps or referring to different subjects

$$N : ({x_{t1}}^{(c_1,m_1)}; {x_{t2}}^{(c_2,m_2)}) \text{ with } t_1 \neq t_2 \quad (2)$$

The goal of this pretraining was to encourage the network to learn modality-invariant and viewpoint-invariant representations that can be exploited for downstream tasks. Further pretraining considerations are reported in the supplementary materials.

4.3. Sequence Length and Framerate Impact

Given the different temporal dependencies of the two proposed tasks, we analyzed the impact of frame rate and time window length on the classification performance of the various networks. Figure 3a illustrates how input length affects network accuracy. The results indicate that gesture recognition is largely time-invariant, with accuracy remaining stable across windows of different lengths. In contrast, AQA benefits significantly from longer temporal contexts, as extended sequences provide more information on execution speed and consistency. Figure 3b presents the effect of temporal subsampling on classification accuracy. While lower frame rates allow for the analysis of longer time windows, they appear to degrade performance for gesture recognition. This suggests that key temporal features are already captured at higher sampling rates, and further extending the time window through undersampling does not add meaningful discriminative information. In contrast, for quality classification, ResNet and ResNeXt show a slight performance improvement when data is downsampled to 7.5 fps.

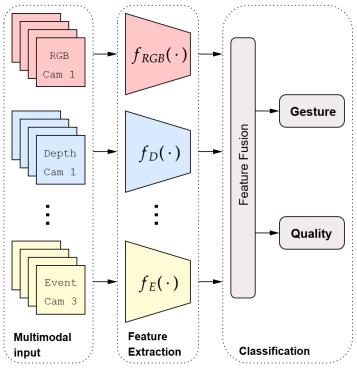


Figure 2. Proposed architecture for multimodal video analysis

4.4. Trigger detection

For trigger detection, we used a baseline pipeline that leverages hand tracking via `mediapipe-hands`. The processing for each recording consisted of:

1. Extracting hand landmarks using `mediapipe-hands` from RGB data, analyzing the `Main` and `Sub` camera separately.
2. Computing reference gesture trajectories for each gesture.
3. Merging reference trajectories from both views using the arithmetic mean and applying a 1D convolution with a fixed-size smoothing window to refine the results.
4. Identifying local extrema in the resulting trajectory, which correspond to the triggers of interest.

For finger tapping, we used the distance between the thumb tip and index tip; for hand opening and closing, the distance between the middle finger and wrist; for the finger-to-nose gesture, the trajectory of the index tip across the image; and for pronation-supination, the distance between the index tip and pinkie tip. We experimented with different smoothing window sizes (3, 5, 7) to assess variations across execution speeds.

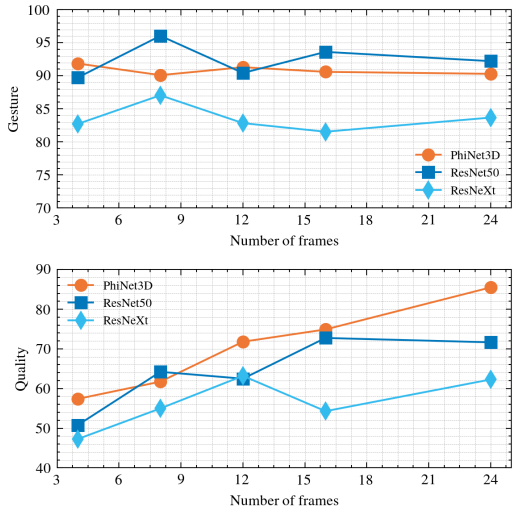
To evaluate baseline performance, we used the mean absolute error (MAE) to measure temporal delay, detection accuracy to quantify the number of correctly identified triggers compared to ground truth, and the false detection ratio (FDR) to account for spurious triggers generated by the detection pipeline.

5. Results

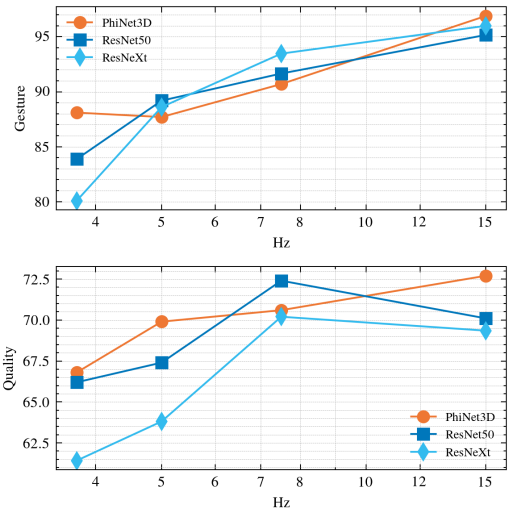
5.1. Gesture classification and action quality

To assess the impact of multimodal inputs on performance, the different architectures were trained on both tasks using an increasing number of input modalities. The results, summarized in Figure 4, illustrate performance trends across different architectures and input configurations.

Interestingly, model performance does not show a strong correlation with the complexity or size of the feature extractors used; even relatively lightweight architectures achieve



(a) Impact of the number of window sizes on network accuracy.



(b) Impact of input framerate on network accuracy.

Figure 3. Comparison of the impact of input length and framerate on network accuracy for the two tasks. (a) Quality estimation shows a strong dependency on window length, while gesture classification is almost independent. (b) Higher framerates are beneficial for gesture classification, although leading to increased computational complexity. Quality estimation benefits slightly from temporal downsampling.

competitive classification results in both tasks. Across all architectures, unimodal inputs yield similar performance, with depth information providing a slight improvement over RGB data from the same camera. Likewise, event-based data perform comparably to other single-modality inputs.

Fusing RGB and depth from a single camera results in a modest performance boost ($+0.1\%$ for gesture recognition and $+1.6\%$ for quality estimation on average). In contrast, incorporating data from two distinct cameras leads to a more substantial improvement ($+1.2\%$ for gesture recognition and $+3.2\%$ for quality estimation on average). Ultimately, leveraging all three modalities in combination yields the highest performance gains across both tasks, with an average accuracy increase of $+3.3\%$ for gesture detection and $+4.5\%$ for quality estimation.

5.2. Trigger detection

Results for trigger detection on the five validation subjects are presented in Table 2. For each task, we report only the result achieved using the optimal smoothing window (according to accuracy). The findings indicate that detecting gesture triggers is generally straightforward, with high detection accuracy across all gestures, despite the simplicity of the baseline approach. However, the main challenge lies in precisely identifying the exact timing of the trigger event, as both MAE and FDR exhibit high values with large standard deviations. Consequently, EHWGesture may serve as a benchmark for trigger detection, particularly for gestures such as pronation-supination and finger-to-nose reaching.

We also examine the relationship between execution speed and smoothing window size. As shown in Figure 5, trigger detection for slow movements significantly benefits from longer smoothing windows, which help reduce false detections. In contrast, FAST and NORMAL gestures achieve comparable performance with smoothing windows of size 5 or 7. This occurs because slower gestures are more ambiguous to segment, as their minima points often correspond to extended plateaus where multiple consecutive frames may be misinterpreted as trigger events. Therefore, integrating AQA predictions related to execution speed into this baseline could enable automatic tuning of smoothing window sizes for the triggering task.

Table 2. Trigger Detection performance (Best Smoothing Window *sw*). Tasks: FT-finger taping, OC-hand opening-closing, PS - Pronation-Supination, NOSE-finger-to-nose.

Task	MAE (s)	Acc. (%)	FDR (%)
FT (sw=3)	0.12	97.37 ± 3.54	16.11 ± 15.13
OC (sw=3)	0.11	98.80 ± 1.78	20.62 ± 18.87
PS (sw=5)	0.27	97.26 ± 5.16	7.98 ± 8.70
NOSE (sw=3)	0.48	98.07 ± 3.08	28.97 ± 19.52

6. Discussion

The results obtained in the experiments demonstrate that EHWGesture can serve as a benchmark for multimodal clinical gesture understanding, supporting diverse and relevant tasks such as gesture classification, AQA, and trigger detection. These two tasks in particular still offer margin for improvement and are pivotal for supporting the creation of more advanced models for AQA to use with real patho-

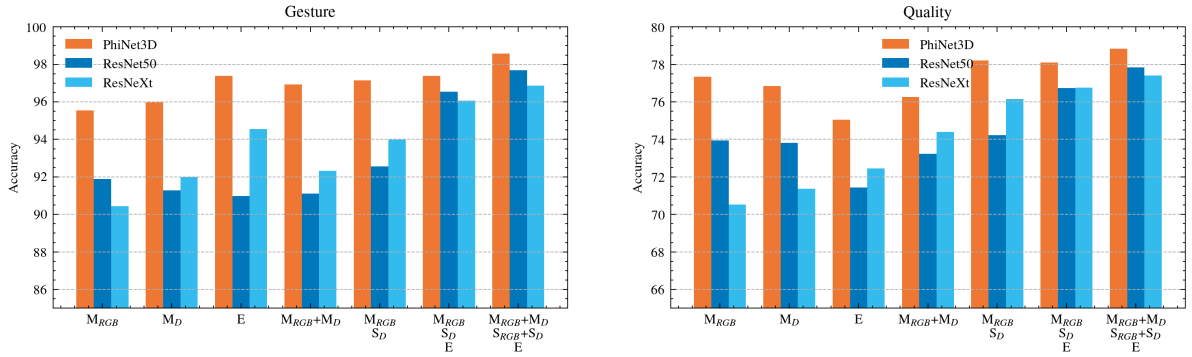


Figure 4. Performance of the tested architectures across both tasks. Increasing the number of input modalities generally improves performance, particularly for larger networks. M: main camera, S: sub camera, E: event, D: depth

logical data, as done by Dadashzadeh *et al*. in [9]. Multimodality, the main contribution of this work, proved essential for improving classification models, and its impact goes beyond the specific gestures of this dataset. EHWGesture can facilitate further research on combining either the three provided raw modalities or additional derived modalities. Moreover, the two Kinect devices and the event camera captured gestures from three distinct perspectives, influencing model training. This aspect opens possibilities for exploring the impact of different viewpoints on gesture understanding.

Ethical considerations: Data collection adhered to GDPR, and all subjects provided written consent to share their data for non-commercial, research-only purposes. Videos are publicly released in an anonymized format, with faces blurred in all frames. We used SAM2 to detect overlapping regions of hands and faces, applying blurring only to the non-intersecting portion. This approach ensures that identifying information cannot be misused for purposes such as generating deepfakes, or facilitating identity theft. Subjects' identities were not stored, as each participant was assigned an anonymous identifier during data collection.

Biases: As this dataset prioritizes providing a comprehensive multimodal data source, subject diversity was somewhat limited. The small number of volunteers does not currently allow for extensive stratification, with only five subjects not conforming to the Caucasian phenotype.

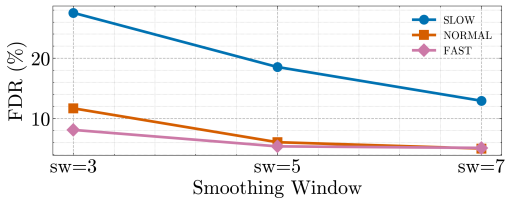


Figure 5. Percentage of wrong trigger detections over total detections considering different smoothing windows. Slower movements improve the most from longer windows.

However, we plan to extend the dataset to improve subject diversity in future iterations, possibly including also pathological subjects.

Limitations: The use of a motion capture system implied recording all trials in the same environment. This limits background diversity in RGB frames, though other modalities remain unaffected. This limitation is, however, coherent with clinical gesture assessment, since examinations are often conducted in standardized settings. In addition, event frames may be influenced by different lighting conditions; while illumination levels were annotated during recordings, extreme lighting conditions were not tested, as they could severely impact reliability of motion capture data. As previously noted, the dataset is biased toward fair skin tones, which may affect the robustness of models trained solely on the RGB modality. However, the multimodal nature of the dataset may help mitigate this limitation. Finally, the baseline trigger detection method does not incorporate a training phase but instead relies on pretrained models and deterministic signal processing. As a result, it may be fragile and prone to inconsistencies when applied to real-world scenarios. Nonetheless, we observed that its performance could be improved by integrating AQA information estimated by the trained multimodal models.

7. Conclusion

This work introduced EHWGesture, a large-scale benchmark for multimodal gesture understanding. The dataset includes gestures from clinical hand dexterity assessment and may support the development of automated models aimed at this application. Additionally, EHWGesture is the first gesture dataset to simultaneously integrate RGB, depth, and event data captured from three different viewpoints while also incorporating an AQA based on gesture execution speed. Baseline experiments demonstrated the dataset's potential and provided insights for future research using this resource.

Acknowledgments

This study is supported by SERICS project (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

References

[1] Muhammad Aitsam, Sergio Davies, and Alessandro Di Nuovo. Event camera-based real-time gesture recognition for improved robotic guidance. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2024. 3

[2] Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, Jeff Kusnitz, Michael Debole, Steve Esser, Tobi Delbruck, Myron Flickner, and Dharmendra Modha. A low power, fully event-based gesture recognition system. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7388–7397, 2017. 3

[3] Gianluca Amprimo, Giulia Masi, Gabriella Olmo, and Claudia Ferraris. Deep learning for hand tracking in parkinson's disease video-based assessment: Current and future perspectives. *Artificial Intelligence in Medicine*, 154:102914, 2024. 2

[4] Gianluca Amprimo, Giulia Masi, Giuseppe Pettiti, Gabriella Olmo, Lorenzo Priano, and Claudia Ferraris. Hand tracking for clinical applications: Validation of the google mediapipe hand (gmh) and the depth-enhanced gmh-d frameworks. *Biomedical Signal Processing and Control*, 96: 106508, 2024. 5

[5] Gibran Benitez-Garcia, Jesus Olivares-Mercado, Gabriel Sanchez-Perez, and Keiji Yanai. Ipn hand: A video dataset and benchmark for real-time continuous hand gesture recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4340–4347, 2021. 3

[6] Y Bi, A Chadha, A Abbas, , E Bourtsoulatze, and Y Andreopoulos. Graph-based object classification for neuromorphic vision sensing. In *2019 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2019. 2

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 5

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020. 6

[9] Amirhossein Dadashzadeh, Shuchao Duan, Alan Whone, and Majid Mirmehdi. Pecop: Parameter efficient continual pretraining for action quality assessment. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 42–52, 2024. 1, 2, 3, 4, 8

[10] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 409–419, 2018. 3

[11] Rohan Ghosh, Anupam Gupta, Andrei Nakagawa, Alcimar Soares, and Nitish Thakor. Spatiotemporal filtering for event-based action recognition. *arXiv preprint arXiv:1903.07067*, 2019. 3

[12] Christopher G Goetz, Barbara C Tilley, Stephanie R Shaftman, Glenn T Stebbins, Stanley Fahn, Pablo Martinez-Martin, Werner Poewe, Cristina Sampaio, Matthew B Stern, Richard Dodel, et al. Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): scale presentation and clinimetric testing results. *Movement disorders: official journal of the Movement Disorder Society*, 23(15):2129–2170, 2008. 4

[13] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. 5

[14] Rahul Jain, Ram Kumar Karsh, and Abul Abbas Barbhuiya. Literature review of vision-based dynamic gesture recognition using deep learning techniques. *Concurrency and Computation: Practice and Experience*, 34(22):e7159, 2022. 1

[15] Alexander Kapitanov, Karina Kvanchiani, Alexander Nagaev, Roman Kraynov, and Andrei Makhliarchuk. Hagrid – hand gesture recognition image dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4572–4581, 2024. 1, 5

[16] Okan Kopuklu, Neslihan Kose, Ahmet Gunduz, and Gerhard Rigoll. Resource efficient 3d convolutional neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 5

[17] Jean-Matthieu Maro, Sio-Hoi Ieng, and Ryad Benosman. Event-based gesture recognition with dynamic background suppression using smartphone computational capabilities. *Frontiers in Neuroscience*, 14, 2020. 3, 5

[18] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. The jester dataset: A large-scale video dataset of human gestures. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 2874–2882, 2019. 1, 3

[19] Sushmita Mitra and Tinku Acharya. Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(3):311–324, 2007. 1

[20] Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4207–4215, 2016. 3

[21] Anton Nuzhdin, Alexander Nagaev, Alexander Sautin, Alexander Kapitanov, and Karina Kvanchiani. Hagridv2: 1m images for static and dynamic hand gesture recognition, 2024. 1

[22] Francesco Paissan, Alberto Ancilotto, and Elisabetta Farella. Phinets: A scalable backbone for low-power ai at the edge. *ACM Trans. Embed. Comput. Syst.*, 21(5), 2022. 5

[23] D. Roalf et al. Quantitative assessment of finger tapping characteristics in mild cognitive impairment, alzheimer's disease, and parkinson's disease. *Journal of Neurology*, 2018. 4

[24] Zinah Raad Saeed, Zurinahni Binti Zainol, BB Zaidan, and Abdullah Hussein Alamoodi. A systematic review on systems-based sensory gloves for sign language pattern recognition: An update from 2017 to 2022. *IEEE Access*, 10:123358–123377, 2022. 1

[25] Sakshi Sharma and Sukhwinder Singh. Vision-based hand gesture recognition using deep learning for the interpretation of sign language. *Expert Systems with Applications*, 182: 115657, 2021. 1

[26] Yuanyuan SHI, Yunan LI, Xiaolong FU, MIAO Kaibin, and MIAO Qiguang. Review of dynamic gesture recognition. *Virtual Reality & Intelligent Hardware*, 3(3):183–206, 2021. Hand and gesture. 1

[27] Jun Wan, Stan Z. Li, Yibing Zhao, Shuai Zhou, Isabelle Guyon, and Sergio Escalera. Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 761–769, 2016. 3

[28] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 5

[29] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. Mediapipe hands: On-device real-time hand tracking. *ArXiv*, abs/2006.10214, 2020. 5

[30] Yifan Zhang, Congqi Cao, Jian Cheng, and Hanqing Lu. Egogesture: A new dataset and benchmark for egocentric hand gesture recognition. *IEEE Transactions on Multimedia*, 20(5):1038–1050, 2018. 1, 3

[31] Yuhan Zhang, Lindong Wu, Weihua He, Ziyang Zhang, Chen Yang, Yaoyuan Wang, Ying Wang, Kun Tian, Jianxing Liao, and Ying Yang. An event-driven spatiotemporal domain adaptation method for dvs gesture recognition. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 69 (3):1332–1336, 2021. 3